Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

A note on independence assumption on binding sites in biological sequence analysis[☆]

Johan Lim[a], Kyeong Eun Lee[b,*]

[a] *Department of Statistics, Seoul National University, Republic of Korea*

[b] *Department of Statistics, Kyungpook National University, Daegu, Republic of Korea*

ARTICLE INFO

Article history:
Received 16 October 2009
Accepted 23 April 2010
Available online 26 May 2010

AMS 2000 subject classifications:
primary 62F99
secondary 62P10

Keywords:
Independence
Markovian dependency
Maximum a posteriori estimate
Sequence analysis

ABSTRACT

Finding significant patterns from sequence data is an important issue in various fields extending well beyond biology. Despite their potential utility, a difficulty in direct application of the recent methods stems from the independence assumption between sites. Markovian assumptions have recently been made for either informative binding sites or non-informative background sites. In this article, it is shown that under certain specifiable conditions, making the independence assumption yields better results in finding binding sites from sequence data with substantial Markovian dependency.

© 2010 The Korean Statistical Society. Published by Elsevier B.V. All rights reserved.

1. Introduction

The utility of sequence data analysis extends well beyond biology to other academic disciplines ranging from computer science to communication to marketing research. For instance, in human computer interaction, Draper (1984) has shown that the users display a great number of variations in their computer use. Some of these variations are from the users' command knowledge and some others are from an analogous variation in the jobs and tasks performed by users. The goal in this area is to develop a system which actively cooperates with its users; such systems are able to initiate interactions with the user as well as to react to users' commands. Also this system can help the user to achieve their tasks by supporting habitual use of the system. Therefore, more efficient algorithms for finding the users' habitual patterns from the sequential data on their computer use are required in this field.

Early models in computational biology including the block motif models and hidden Markov models use the product multinomial model to transform the multiple alignment problems into an issue of statistical inferences (Baldi, Chauvin, McClure, Myers, & Hunkapiller, 1994). The block motif model models the binding sites in biological sequences as un-gapped Markovian blocks (Lawrence & Reilly, 1990; Lawrence et al., 1993; Liu, 1994; Liu, Neuwald, & Lawrence, 1995). The hidden Markov model assumes the observed informative sequences as though they were generated by the hypothetical ancestral model via mutation (Baldi et al., 1994; Eddy, 1995; Krogh, Brown, Mian, Sjölander, & Haussler, 1994). Thus, it allows random gaps within an informative block. Here, each site has its single observation, and the site is "informative" implying that it has information on particular biological traits.

[☆] This research was supported by Kyungpook National University Research Fund, 2009.

[*] Corresponding author. Tel.: +82 53 950 7358.
E-mail address: artlee@knu.ac.kr (K.E. Lee).

1226-3192/$ – see front matter © 2010 The Korean Statistical Society. Published by Elsevier B.V. All rights reserved.

doi:10.1016/j.jkss.2010.04.005

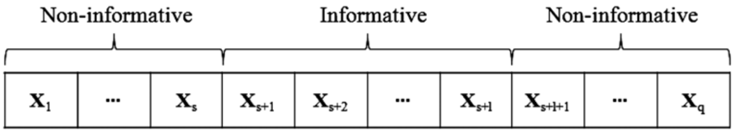

Fig. 1. Schematic plot for background and binding sites; "non-informative" implies background sites and "informative" implies binding sites.

The existing models and statistical procedures assume independence among observations from different sites. In practice, however, this assumption could be disparate from real biological sequences. For this, recently, much effort has been made to model the dependency between sites. Ellrott, Yang, Sladek, and Jiang (2002) assume Markovian binding sites and independent background sites. On the other hand, Kim, Tharakaraman, and Spouge (2006); Kim, Tharakaraman, Marino-ramirez, and Spouge (2008) and Thijs et al. (2001) use models of independent binding sites and Markovian background sites.

In this paper, we ask "what if we apply a simple independent model to dependent biological sequences?". We show that, under certain specifiable conditions, making the independence assumption yields better results in analyzing sequence data with substantial dependency among observations from different sites. By doing so, we attempt to show the utility of the independent model in analyzing various types of sequence data. We assume Markovian binding sites and independent backgrounds as in Ellrott et al. (2002) in this paper, but the results are also applied to the models with independent binding sites and Markov backgrounds by Kim et al. (2006, 2008) and Thijs et al. (2001). We obtain a sufficient condition under which the fake independence assumption works better than the true Markovian dependence assumption. We illustrate this using a simple two state sequence data example. The example provides an explanation as to why the independent model works better than the true dependent model.

The paper is below is as follows. We explain our main results in Section 2. We provide an illustrative example in Section 3. Section 4 proofs of main results.

2. Main results

Sequence data is composed of observations from either binding sites or background sites, where an observation in each site has at most r possible values representing different states of the site. Let $s+1$ and l be the unknown starting site and the length of binding sites, respectively. Also, we let q be the total length of the observed sequence including the binding sites. We further assume that the observations from binding sites are dependent on each other. Suppose the observed sequence is

$$\underline{\mathbf{X}} = \left(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_q\right) = \begin{pmatrix} X_{11} & X_{21} & \cdots & X_{q1} \\ X_{12} & X_{22} & \cdots & X_{q2} \\ \vdots & \vdots & \cdots & \vdots \\ X_{1r} & X_{2r} & \cdots & X_{qr} \end{pmatrix},$$

where $X_{ij} = 1$ if the ith site has state j, and, otherwise, 0 (Fig. 1).

We define the joint probabilities of the observed sequence $P^{\mathrm{I}}(\underline{\mathbf{X}}|\Delta)$ and $P^{\mathrm{D}}(\underline{\mathbf{X}}|\Delta)$ under independent and dependent assumption on binding sites running from $\Delta+1$, respectively:

$$P^{\mathrm{I}}(\underline{\mathbf{X}}|\Delta) \equiv \prod_{m=1}^{l} P^{\mathrm{D}}\left(\mathbf{X}_{\Delta+m}\right) \prod_{k \notin \{(\Delta+1):(\Delta+l)\}} P_0(\mathbf{X}_k)$$

$$P^{\mathrm{D}}(\underline{\mathbf{X}}|\Delta) \equiv P^{\mathrm{D}}\left(\mathbf{X}_{(\Delta+1):(\Delta+l)}\right) \prod_{k \notin \{(\Delta+1):(\Delta+l)\}} P_0(\mathbf{X}_k),$$

where $(a : b)$ implies the sequence $a, a+1, \ldots, b$, Δ is the assumed starting position of the binding sites, $P_0(\cdot)$ is the probability measure of background sites, and $P^{\mathrm{D}}(\cdot)$ is the probability measure of binding sites under the true dependence assumption. Hereafter, s is the true starting position of the binding sites and is assumed to be 0 without loss of generality.

In this paper, we assume that the observations on binding sites are from a first order stationary Markov chain (MC). We conjecture that the results of the paper are not limited to the first order MC, but also true for a higher order MC. The higher order MC has a more complicated transition probability matrix **P** than the first order MC in this paper. However, all computations below are applicable to this new transition matrix regardless of its complexity.

A main objective of sequence analysis is to find the location of unknown binding sites (equivalently, s which is 0) and estimate its probabilities. To do this, under the block motif model or HMM, the maximum a posteriori estimate (MAP), that is

$$\widehat{s}^{\mathrm{I}} = \operatorname{argmax}_{\Delta} P^{\mathrm{I}}\left(\Delta|\underline{\mathbf{X}}\right)$$

and

$$\widehat{s}^{\mathrm{D}} = \operatorname{argmax}_{\Delta} P^{\mathrm{D}}\left(\Delta|\underline{\mathbf{X}}\right),$$

is commonly used.

Download English Version:

https://daneshyari.com/en/article/1144913

Download Persian Version:

https://daneshyari.com/article/1144913

Daneshyari.com